GEA1000 Cheatsheet

github.com/reidenong/cheatsheets, AY23/24 S2

Chapter 1: Data Collection & Sampling

Definitions

- **Population**: A group of individuals or items to study.
- **Population parameter**: A numerical fact of a population.
- **Sample**: A subset of a population.
- **Sample statistic**: A numerical fact of a sample.
- **Estimate**: An inference of a parameter based on a sample.
- **Sampling frame**: A list from which a sample is drawn.
- **Census**: A study of an entire population.
- **Bias**
  - **Selection bias**: Biased selection of units into the sample, caused by imperfect sampling frames or non-probability sampling.
  - **Non-response bias**: Systematic exclusion of non-respondents, due to reasons like inconvenience or unwillingness.

Probability Sampling

- **Simple random sampling**: Each member has an equal chance of being selected.
- **Systematic sampling**: Select every kth member, may be biased given biased list.
- **Stratified sampling**: Divide population into strata of similar characteristics, then sample from each stratum.
- **Cluster sampling**: Divide population into clusters, then sample all members in selected clusters.
- **Random Assignment**: Impartially allocate subjects into test and control groups.

Generalisation can only be done with probability sampling.

Non-Probability Sampling

1. **Convenience Sampling**: Sampling based on availability.
2. **Volunteer Sampling**

Summary

Plan	Pros	Cons
Simple Random Sampling	Good representation of population	Time-consuming, information may not be accesible
Systematic Sampling	Simple process	Potentially under-representing population
Stratified Sampling	Good representation of the sample by stratum	Requires sampling frame and criteria for classification of the population into stratum
Cluster Sampling	Less time-consuming, cheap	Clusters must be heterogenous and not have characteristics

Variables

Definitions

- **Variable**: A attribute that can be measured or labelled
- **Data Set**: A collection of data
- **Independent Variables**: Variables that are manipulated
- **Dependent Variables**: Variables that are measured
- **Categorical variables**: Categorized or labelled
  - **Ordinal**: Has a natural order
  - **Nominal**: No natural order
- **Numerical variables**: Numerical values
  - **Discrete**: Countable
  - **Continuous**: Uncountable

Summary Statistics

Averages

- Mean, Median, Mode, Proportions

Spread

Sample Variance, Var =

$$\frac{1}{n-1}\sum (x_i - \bar{x})^2$$

SD, $s_x = \sqrt{\text{Var}}$ is in the same units as x.

Coefficient of Variance $= \frac{s_x}{\bar{x}}$ has no units. It compares the degree of variation across different variables within a data set.

Median with IQR is preferred when distribution of data is not symmetrical or if there are outliers.

Study Design

Experimental Study / Controlled Experiment

Independent variable is intentionally manipulated to observe the effect on a Dependent variable. Provides evidence for cause-and-effect relationship

Observational Study

Observes individuals and measures variables of interest without influencing the outcome. Cannot establish cause-and-effect relationship.

Fallacies

Ecological fallacy occurs when we use aggregate level (ecological) correlation to draw conclusions about individual data.

Atomistic fallacy occurs when we use individual level correlation to draw conclusions about group data.

Chapter 2. Categorical Data Analysis

Rates

When there is no association between A and B, we have rate $(A \mid B) =$ rate $(A \mid B')$

When there is positive association between A and B, we have

$$\text{rate } (A \mid B) > \text{rate } (A \mid B')$$
$$\text{rate } (B \mid A) > \text{rate } (B \mid A')$$
$$\text{rate } (A' \mid B') > \text{rate } (A' \mid B)$$
$$\text{rate } (B' \mid A') > \text{rate } (B' \mid A)$$

When there is negative association between A and B,

$$\text{rate } (A \mid B) < \text{rate } (A \mid B')$$
$$\text{rate } (B \mid A) < \text{rate } (B \mid A')$$
$$\text{rate } (A' \mid B') < \text{rate } (A' \mid B)$$
$$\text{rate } (B' \mid A') < \text{rate } (B' \mid A)$$

Symmetry rules

$$\text{rate } (A \mid B) > \text{rate } (A \mid B') \Leftrightarrow \text{rate } (B \mid A) > \text{rate } (B \mid A')$$
$$\text{rate } (A \mid B) < \text{rate } (A \mid B') \Leftrightarrow \text{rate } (B \mid A) < \text{rate } (B \mid A')$$
$$\text{rate } (A \mid B) = \text{rate } (A \mid B') \Leftrightarrow \text{rate } (B \mid A) = \text{rate } (B \mid A')$$

Properties of Rates

1. rate (A) is always between rate $(A \mid B)$ and rate $(A \mid B')$
2. The closer rate(B) is to 100%, the closer rate(A|B) is to rate(A)
3. If rate(B) = 50%, rate(A) = $\frac{1}{2}$[rate $(A \mid B') +$ rate $(A \mid B)$]
4. If rate(A | B) = rate(A | B'), then rate(A) = rate(A | B) = rate(A | B')

Simpson's Paradox

Simpson's Paradox occurs when a trend appears in more than half the groups of data but disappears or reverses when the groups are combined. Disappear meaning the two variables are no longer associated.

Simpson's paradox directly implies the existence of a confounding variable.

Confounder

A third variable that is associated with both the the independent and dependent variables

Sensitivity and Specificity

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

Chapter 3. Numerical Data Analysis

Univariate EDA

Bell distributions

- Symmetrical data
  - Mean = Median = Mode
  - 68% of data within 1 SD
  - 95% of data within 2 SD
- Left Skewed (left tail, right peak)
  - Mean < Median < Mode
- Right Skewed (right tail, left peak)
  - Mean > Median > Mode

Boxplots

Minimum, Q_1, Median, Q_3, Maximum

1. Find Median
2. Find Median of lower half (Q_1) and upper half (Q_3)

$$IQR = Q_3 - Q_1$$

Outliers

A data point x is an outlier if

$$x > Q_3 + 1.5 \times IQR$$
$$x < Q_1 - 1.5 \times IQR$$

Boxplot Analysis

Shape

- Symmetrical: Q_1 and Q_3 are equidistant to median
- Right Skewed: Q_1 is closer to the median
- Left Skewed: Q_3 is closer to the median

Spread

- IQR gives a idea of spread for the middle 50%

Histogram vs Boxplot

Histogram

- Shows shape/distribution of data
- Better when there are greater frequencies
- Number of data points are better shown

Boxplot

- Can compare distributions of different data sets by putting boxplots side by side
- Better at identifying and indicating outliers
- Note that two distributions with a very different number of data points may have the same boxplot

Bivariate EDA

Deterministic means the value of one variable can be exactly determined by the value of another variable.

Direction

- Positive relationship occurs when a increase in one variable leads to an increase in another (+ve gradient)
- Negative relationship occurs when a increase in one variable leads to a decrease in another (-ve gradient)
- Strength of the relationship is determined by how closely the data points follow the trend line

Correlation Coefficient, r

Summarises the direction and strength of linear association between two variables.

- $-1 \leq r \leq 1$

Sign

- $r > 0 \Rightarrow$ positive association
- $r < 0 \Rightarrow$ negative association
- $r = 0 \Rightarrow$ no *linear* association
- $r = 1 \Rightarrow$ perfect positive association

Magnitude

- $0 < |r| < 0.3 \Rightarrow$ weak association
- $0.3 < |r| < 0.7 \Rightarrow$ moderate association
- $0.7 < |r| < 1 \Rightarrow$ strong association

Computing r

$$r = \frac{\text{Population Covariance}}{\text{Population SD of x} \times \text{Population SD of y}}$$
$$= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}}$$

* r is not affected by interchanging the x and y variables.
** r is not affected by adding/multiplying a constant to all values of x or y.

Linear Regression

If we believe two variables are linearly associated, we can model the relationship using a straight line to predict the *average* value of the dependent variable given the independent variable.

- The least square regression line always passes through the point $(\bar{x}, \bar{y})$.
- A regression line used to predict y using x cannot be used to predict x using y, as it is modelled to minimise squared errors in the y direction.
- For regression line $Y = mX + b$,

$$m = \left(\frac{s_y}{s_x}\right) r$$

where s_x / s_y refers to the standard deviations of X/Y. It then follows that $r > 0 \Rightarrow m > 0$, and vice versa.

Chapter 4. Statistical Inference

Available data can be used to make inferences about the population if the data can be considered to be representative with regards to the question of interest.

Sample statistic = population parameter + bias + random error

Confidence Interval

A range of values that is likely to contain the estimate of a population parameter with a certain level of confidence.

$$CI = \text{point estimate} \pm \text{margin of error}$$

For a 95% CI, we are 95% confident that the interval contains the population parameter.

CI for Population Proportions

Given a sample proportion p^* and sample size n, the CI for population proportion is

$$p^* \pm z^* \times \sqrt{\frac{p^*(1-p^*)}{n}}$$
$$\text{if CI is 90\%, } z^* = 1.645$$
$$\text{if CI is 95\%, } z^* = 1.96$$

CI for Population Mean

Given a sample with mean $\bar{x}$, sample standard deviation s and sample size n, the CI for population mean is

$$\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$$

where t^* depends on the sample size n and the desired confidence level.

Hypothesis Testing

p-value

- The probability of obtaining a result as extreme or more extreme as the observed result, assuming the null hypothesis is true.
- p-value $<$ than the significance level α
  - We reject the null hypothesis in favour of the alternative hypothesis.
- p-value $\geq$ than the significance level α
  - We fail to reject the null hypothesis.
  - Does not mean the null hypothesis is true

Types of hypothesis testing

- Hypothesis testing for population proportion
- Hypothesis testing for population mean
- Hypothesis testing for association
  - H_0 : There is no association
  - H_1 : There is an association
  - Chi-Square test is used

Fallacies

Conjunction Fallacy

- The probability of two events occurring together is always less than the probability of either event occurring alone.

Base Rate Fallacy

- Ignoring the base rate of an event when calculating the probability of that event.